Bagian ini berisi tahapan pemrosesan bahasa alami yang akan diterapkan. Gambar 1 menunjukkan diagram alir proses yang akan dilakukan untuk analisis sentimen pada lirik lagu.

Data Cleaning

Tokenization

Case Folding

Lemmatization

Stemming

Stopwords Removal

Data Collection

Data Preprocessing

Feature and Label Extraction (TF IDF / Count Vectorizer)

Feature Selection

Modelling with SVM

Evaluation and Result

**Gambar 1. Tahapan Proses Analisis Sentimen pada Lirik Lagu**

## ***Data Collection***

*Dataset* yang digunakan pada proyek ini merupakan *dataset* lirik lagu dalam bahasa inggris yang terdiri dari 400 *music clips* dan telah diberikan label untuk setiap kategori dari lirik lagu tersebut, apakah termasuk kategori *angry*, *happy*, *relaxed*, dan *sad*. *Dataset* tersebut merupakan data yang bersifat statis karena dalam format file CSV (*Comma Separated Values*).

## ***Data Preprocessing***

*Data Preprocessing* merupakan tahap pemrosesan awal yang dilakukan sebelum teks diolah ke tahap selanjutnya. Data yang diperoleh masih dalam format yang tidak terstruktur, dimana masih terdapat *noise* pada dataset tersebut. Selain itu, data masih dalam format *raw data* sehingga tidak memungkinkan untuk melakukan analisis pada *raw data*. Oleh karena itu perlu dilakukan *data preprocessing* untuk menghilagkan kata-kata pada teks atau dokumen yang mengandung beberapa format yang keberadaannya tidak penting dalam *text mining*.

### ***Data Cleaning***

Data yang diperoleh dari dataset memiliki beberapa *noise* yang perlu dibersihkan, misalnya string yang kosong (*incomplete data*), salah penulisan pada kata (*typo*), tanda baca. Pada tahapan ini, dilakukan pembersihan data dari karakter-karakter yang tidak diinginkan, misalnya tanda baca, angka, spasi, atau karakter-karakter tertentu pada data tersebut. Berikut adalah beberapa contoh data yang *noise* pada dataset:

* SlavesÃ¯Â¼ÂŒHebrews born to serve, to the pharaoh HeedÃ¯Â¼ÂŒTo his every word, live in fear FaithÃ¯Â¼ÂŒOf the unknown one
* Â·Ã«ÃªÃ˜Ã¦Â¥

Beberapa karakter tersebut perlu dibersihkan. Beberapa cara yang dapat dilakukan untuk membersihkan data adalah sebagai berikut:

* Mengabaikan data yang hilang dengan cara menghapus data tersebut, namun cara ini tidak cukup efektif ketika terdapat banyak data yang hilang.
* Mengisi *missing value*, namun cara ini tidak cukup efektif ketika datanya besar dan membutuhkan waktu yang lama.
* Memperbaiki penulisan kata yang kurang tepat.
* Menghapus karakter-karakter yang bersifat *noisei* pada dataset.

### ***Case Folding***

Setelah data dibersihkan, pada tahapan ini dilakukan proses konversi data teks yang terdapat pada *dataset* menjadi ke dalam bentuk *lower case*. Hal ini untuk memudahkan preproses data di tahap selanjutnya. Berikut adalah contoh *case* *folding*:

**Tabel 1. Case Folding**

|  |  |
| --- | --- |
| ***Text*** | ***Case* *Folding*** |
| LONG time ago, there lived a King | long time ago, there lived a king |

### ***Stopwords Removal***

Pada tahapan ini akan dilakukan proses untuk menghilangkan *stopwords*. Kata-kata seperti “this”, “there”, “the”, “a”, “of” merupakan *stopwords* yang biasa digunakan dalam teks. Jika dibandingkan dengan kata lainnya, frekuensi kemunculan *stopwords* lebih banyak. *Stopwords* merupakan elemen penting yang perlu untuk dihilangkan karena membuat teks yang diproses menjadi lebih banyak. Selain itu, *stopwords* juga kurang dibutuhkan dalam proses *text* *mining*. Sehingga, menghilangkan *stopwords* akan mengurangi dimensi dari data dan akan meningkatkan kecepatan dalam memproses data. Contoh *stopwords removal* adalah sebagai berikut.

**Tabel 2. Stopwords Removal**

|  |  |
| --- | --- |
| ***Text*** | ***Stopwords Removal*** |
| *LONG time ago, there lived a King* | *LONG time ago, lived King* |

### ***Tokenization***

Tujuan dari tahapan ini adalah untuk memecah teks menjadi unit yang relevan, seperti *words,* simbol, *phrases*, atau elemen bermakna lainnya yang disebut juga dengan *token*. Contoh *tokenization* adalah sebagai berikut.

**Tabel 3. Tokenization**

|  |  |
| --- | --- |
| ***Text*** | ***Tokenization*** |
| *LONG time ago, there lived a King* | *long, time, ago, there, lived, a, king* |

### ***Stemming***

*Stemming* merupakan sebuah proses menghilangkan imbuhan pada kata untuk mendapatkan kata dasar atau *stem* dari kata tersebut. Contohnya, kata *develop, developer, development, developing* akan dipetakan ke kata dasar develop. Tujuan dari tahapan ini adalah untuk menghilangkan variasi imbuhan yang mungkin pada suatu kata, sehingga akan mengurangi jumlah kata yang diidentifikasi. Contoh *stemming* adalah sebagai berikut:

*develop*

*develop*

*Stemming*

*developer*

*development*

*developing*

**Gambar 2. Stemming**

### ***Lemmatization***

*Lemmatization* merupakan sebuah proses untuk mengubah kata menjadi bentuk dasarnya sesuai dengan kata-kata yang terdapat dalam kamus. Contoh *lemmatization* adalah sebagai berikut.

**Tabel 4. Lemmatization**

|  |  |
| --- | --- |
| ***Word*** | ***Word Lemmatization*** |
| *Good* | *Good* |
| *Better* | *Good* |
| *Best* | *Good* |
| *Leafs* | *Leaf* |
| *Leaves* | *Leaf* |